LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-642881

Towards fully automated high performance computing drug discovery: A massively parallel virtual screening pipeline for docking and MM/GBSA rescoring to improve enrichment

X. Zhang, S. E. Wong, F. C. Lightstone

August 27, 2013

Journal of Chemical Information and Modeling

Disclaimer

Towards fully automated high performance computing drug discovery: A massively parallel virtual screening pipeline for docking and MM/GBSA rescoring to improve enrichment

Xiaohua Zhang, Sergio E. Wong, and Felice C. Lightstone*

Biosciences & Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Lab, Livermore, CA 94550.

*To whom correspondence should be addressed: lightstone1@llnl.gov

Abstract

In this work we announce and evaluate a high throughput virtual screening pipeline for *in-silico* screening of virtual compound databases using high performance computing (HPC). Notable features of this pipeline are an automated receptor preparation scheme with unsupervised binding site identification. The pipeline includes receptor/target preparation, ligand preparation, VinaLC docking calculation, and MM/GBSA rescoring using the GB model by Onufriev and co-workers. Furthermore, we leverage HPC resources to perform an unprecedented, comprehensive evaluation of MMGBSA rescoring when applied to the DUD data set (Directory of Useful Decoys), which includes 38 protein targets and a total of ~0.7 million actives and decoys. The computer wall time for virtual screening has been reduced drastically on HPC machines, which increases the feasibility of extremely large ligand database screening with more accurate methods. HPC resources allowed us to rescore 20 poses per compound and evaluate the optimal number of poses to rescore. We find that keeping 5-10 poses is a good compromise between accuracy and computational expense. Overall the results demonstrate that MM/GBSA rescoring has higher average ROC area under curve (AUC) values and consistently better early recovery of actives than Vina docking. On average MM/GBSA rescoring improves the enrichment performance compared to Vina docking. Specifically, the enrichment performance is target-dependent. MM/GBSA rescoring significantly out performs Vina docking for the folate enzymes, kinases and several other enzymes. The more accurate energy function and solvation terms of the MM/GBSA method allow MM/GBSA to achieve better enrichment, but the rescoring is still limited by the docking method to generate the poses with the correct binding modes.

Introduction

Accurately and efficiently predicting the binding affinities of putative protein-ligand complexes is crucial for structure-based drug virtual screening[1,2]. Molecular docking methods with various scoring functions are usually employed to access the binding affinities between compounds and drug targets in the early stage of structure-based virtual screening[3,4]. To achieve high throughput, the scoring functions often use less computationally intensive methods, such as molecular mechanics force-field methods, empirical scoring functions, and/or knowledge-based potentials[5,6]. The scoring functions often simplify the calculation by neglecting important terms that are known to influence the binding affinity, such as, solvation, entropy, receptor flexibility, etc [7,8]. A very popular practice is to rescore top-ranking docking poses using more accurate, albeit computationally costly, methods to overcome shortcomings in the docking scoring function [9-11].

Solvation effects, mainly contributed by water molecules in the biological systems, play a critical role in ligand binding by providing bulk solvent stabilization and solute-desolvation, increasing the entropic contribution with the release of water molecules in the active site upon binding, serving as molecular bridges between the ligand and receptor, etc[12]. There are two main molecular mechanical (MM) models to simulate water: explicit[13] and implicit/continuum solvent models[14]. Explicit water models treat solvent by including individual water molecules, while implicit models represent water as a homogeneous dielectric. The continuum model is much less computationally expensive than the explicit model, which makes it an ideal method to carry out the rescoring of an enormous number of docking poses. Numerically solving the Poisson-

Boltzmann (PB) equation yields the electrostatic potential for a given system. An even faster approach uses the Generalized Born (GB) model, which yields accuracy comparable to the PB method.

Combining molecular mechanics and implicit solvent models, Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) and Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) have been widely applied to calculate the free energy of binding for ligand-receptor complexes[15-17]. Compared to other free energy methods, such as free energy perturbation and thermodynamic integration methods, the MM/PB(GB)SA method is much less computationally intensive. As a standard protocol, the MM/PB(GB)SA method has been implemented in many molecular mechanics simulation packages, e.g. Amber[18,19], CHARMM[20], Gromacs[21], etc. In recent years, many reviews have been published on the development and progress of the implicit solvent model and the MM/PB(GB)SA method[22-28]. Chen et. al. have optimized atomic radii and protein backbone torsional parameters of implicit solvent models to better match potentials of mean force (PMF) calculated from explicit solvent free energy simulations[29]. Explicit solvent molecules have been added to the continuum solvent calculations by Kelly and co-workers for the calculation of aqueous acid dissociation constants[30]. Accurate absolute solvation free energies of small molecules using an implicit solvent model have been calculated with an automated protocol developed at Roux group[31]. A GB model is extended by the molecular volume correction term by Mongan et.al., which largely corrects the solvent-excluded volume of each pair of atoms[32]. A new GB model, developed by Labute, estimates the free energy of hydration using London dispersion instead of atomic surface area[33].

In this work we report two things: 1) the development of an automated pipeline to prepare, dock and re-score protein-ligand complexes and 2) a very large-scale validation of the GB model developed by Onufriev and coworkers[34] for the purposes of enrichment in virtual screening experiments., The accuracy of poses and binding free energies for the Onufriev et al[34] MM-GBSA method was investigated by Hou et. al[35-37]. Here we focus on enrichment of binders versus non-binders using the MMGBSA rescoring. Huang and colleagues[38] have reported on a combination of docking and rescoring of nine protein targets with a modest pool of actives and decoys. However, to the best of our knowledge, there are no reports that systematically study the enrichment factor of the rescoring method on large databases with a variety of targets and thousands to millions of actives and decoys. In this article, we are able to perform such a study by leveraging the high performance computing (HPC) resource at Lawrence Livermore National Laboratory and with our in-house developed fully automated high throughput virtual screening pipeline. The docking and rescoring results of the Directory of Usefully Decoys Enhanced (DUD-E) data set[39], as calculated by the VinaLC docking program[40] and our in-house rescoring protocol, are presented in this study.

Method

Workflow of Docking and Rescoring

The workflow of the fully automated high throughput virtual screening pipeline is shown in Figure 1. The overall workflow is to treat the receptors/targets and ligands by the programs, preReceptors and preLigands, respectively. Then, complexes are generated by docking the ligands into the active sites of receptors/targets using the VinaLC docking program[40]. The top 20 docking poses of each complex are re-scored with MM/GBSA and re-ranked by their calculated binding free energies.

The detailed procedures involved in the workflow are described as follows. For the receptors, the raw PDB files are processed by our in-house program to identify active sites[41,42]. The structures of the receptors are protonated, and the centroids of the active sites/binding sites are determined. These pre-treated receptor structures are used as input for the preReceptor program. The preReceptor program firstly determines the dimensions of docking grids by utilizing the dms[43] and sphgen programs[44]. The dms program calculates the molecular surface of receptor, and the sphgen program fills the active site of receptor with spheres. In order to reduce the computer time, the receptor is cut at a radius of 30 Å from the centroid of active site. The dimensions of docking grids are determined by finding the distribution of spheres along the X-, Y-, and Z-axis. The cutoffs are set when the distribution of spheres changes drastically. The dimensions of the docking grids and centroid of the active site are stored for docking calculations in the next step. The Amber forcefield f99SB[45] is employed in the calculation for the receptor. Any non-standard amino acids are converted to alanine, only if it is not present in the active site. If present in the active site, the non-standard amino acids are pre-

calculated and stored in the library. Energy minimization of the receptor is carried out using the MM/GBSA implemented in the sander program of the Amber package[18]. Energy minimization is separated into two steps: 1) the structures are minimized with heavy atom constraints, and 2) then, all the constraints are removed. The PDB files of energy-minimized receptor structures are converted to PDBQT files, which are used in the docking procedure. During the conversion, the non-polar hydrogen atoms are removed from the receptor structures. Ligands use the Amber GAFF forcefield[46] as determined by the antechamber program[47] in the Amber package. Partial charges of ligands are calculated using the AM1-BCC method[48]. The structures of ligands are energetically minimized by the MM/GBSA method implemented in sander. The atomic radii developed by Onufriev and coworkers (Amber input parameter *igb*=5) are chosen for all GB calculations[34]. Those atoms with GB radii missing from the original program (i.e. fluorine, using a GB radius of 1.47 Å) are added into the Sander program. The PDB files of energy-minimized ligand structures are converted to multiple-structure PDBQT files. As with the receptors, non-polar hydrogen atoms are removed from the ligand structures.

The VinaLC parallel docking program is employed to generated ligand-receptor complexes from a list of receptors and a list of ligands. The docking grid granularity is set to 0.333 Å. The exhaustiveness is set to 12, so that 12 Monte Carlo simulations for searching docking poses are run for each complex. The top 20 docking poses of each ligand are saved for the re-scoring step.

Because the non-polar hydrogen atoms have been removed from the ligands before the docking calculations, they are re-generated using the tleap program in the Amber

package, according to the Amber forcefield, and saved in PDB files in preparation for the MM/GBSA rescoring procedure. The PDB files of the ligand and its corresponding receptor are concatenated into one PDB file, which is the PDB file for the complex. Energy minimization is performed for the complex with the receptor portion constrained and using the MM/GBSA method. The final total energy of the complex, excluding the constraint energy, is extracted. The final total energies of the ligand and receptor are also extracted from the previous calculation. The final total energy includes standard energy terms with two additional terms, GB and surface energies, calculated by the MM/GBSA method. The binding free energy is calculated from the final total energy of the complex, subtracting those of the ligand and the receptor. The top 20 docking poses are re-ranked based on the resulting binding free energy.

Code Implementation

Based on the workflow (Figure 1), four parallel programs, preReceptors, preLigands, VinaLC, and mmgbsa, have been developed to perform the calculations at different steps (Figure 2). VinaLC uses an MPI and multithreading hybrid scheme, which has been previously described[40]. For the other three programs, they have a similar design that contains an MPI framework and provides interfaces to call applications in the Amber package in order to calculate a massive number of ligands, receptors, and complexes in parallel on the HPC machines.

The MPI framework employs a master-slave scheme as illustrated in Figure 2. The master process is shown on the left side, and the slave processes are on the right. The master process is in charge of job dispatching, input/output data handling, job tracking, etc. The slave processes receive the input data from the master, perform the actual

calculation, and send the results/error messages to the master for output. The master process goes through every combination of the receptor and ligand. The master process tries to receive the rank of any free slave process. If there are still jobs in the queue, the master process sends an unfinished job flag to the free slave process. All the input and output data are handled by the master process. The input data are packed into one data package so that only one pair of MPI send/recv calls is required, to reduce the MPI overhead. The output data are treated with the same approach. The master process sends the input data required for the docking or rescoring calculations to the slave process. After receiving the input data, the slave process performs the calculation. The slave process sends the output data back to the master process when it finishes the assigned calculation from the master process. Only after the master process has assigned each slave process with a docking calculation will the master process start collecting the output data. Once the output data from the slave process is collected, then the master process will give that slave process another job. When there are no jobs left in queue, the master process sends a finished job flag to free the slave processes. By implementing such a master-slave MPI scheme, the master is in charge of job dispatching, input, and output while the slave processes are kept busy by running individual calculations until all the calculations are finished. In the slave process, the interfaces to the applications in the Amber package are implemented and are wrapped with try-catch clauses to catch any error that could arise from the calculation. Those errors are passed to the master process and stored in the job-tracking file in XML format. The master node will also track the status of the completed jobs. This information is also saved in the job-tracking XML file.

The program currently supports Linux, IBM BG/Q, and Mac OS operating systems and relies on two external libraries, BOOST and MPI. Typical commands to run the workflow, using preReceptors, preLigands, VinaLC, and mmgbsa programs with a SLURM[49] job schedule, are:

srun -N50 -n800 preReceptors --recList recList.txt [--xml jobtrack]

srun -N80 -n960 preLigands --sdf ligand.sdf [--xml jobtrack]

srun -N1284 -n1284 -c12 vinaLC --recList recList.txt --ligList ligList.txt --geoList geoList.txt [--exhaustiveness 12 --granularity 0.333 --num_modes 20]

srun -N1284 -n15408 mmgbsa --recList recList.txt --ligList ligList.txt [--xml jobtrack],

where the flags in "[]" are optional. The default values are used, if these options are not present. The combination of four commands performs the docking and MM/GBSA rescoring of the ligands in "ligand.sdf" against targets in the "recList.txt".

Benchmark Data Set

The DUD data set is a very popular data set in benchmarking docking programs[40,50]. The original DUD data set has 40 protein targets, 2,950 actives overall, and 36 decoys on average for each active. Recently, an upgrade version data set, DUD-E, has been released by Irwin et. al.[39], with a total of 102 protein targets, 22,886 actives overall, and 50 decoys on average for each active. The new DUD-E data set improves chemotype diversity, decreases net formal charge imbalance between actives and decoys, eliminates the false decoys, etc. In our previous study, we used the DUD data set to benchmark the VinaLC docking program[40]. With the new DUD-E data set available, we selected the

38 protein targets, which are included in both the original and new DUD data set to allow for easy comparison. In the DUD-E data set, Platelet-derived growth factor receptor β was dropped because it is a homology model. Estrogen receptor α (ESR1) is a single target in DUD-E, whereas it was split into agonists and antagonists previously. The new actives and decoys of those 38 protein targets in the DUD-E data set were employed in benchmarking our docking-rescoring pipeline.

At first, both MM/PBSA and MM/GBSA methods were employed in the binding free energy calculations in a preliminary study. However, we found that MM/PBSA does not out-perform MM/GBSA based on several examples calculated in the preliminary study. This finding is in agreement with the studies by Hou and others[35,51], where their results showed that MM/PBSA performed better in calculating absolute, but not necessarily relative, free energies than MM/GBSA. Because MM/PBSA is much more computationally intensive than MM/GBSA, and the data set has a large number of actives and decoys, only the MM/GBSA method is used for rescoring. In some extremely rare cases, some single point MM/GBSA energy minimizations could not converge and are excluded from the final results.

Docking scores and MM/GBSA rescoring values are directly compared head-to-head in this study. Some ligands in the DUD-E data set have multiple structures due to chirality and/or tautomerization. For the MM/GBSA method, only one structure associated with the lowest binding free energy is used in the final results. For the Vina docking method, only one structure associated with the top Vina docking score is used in the final results.

Enrichment Factor

There are many ways to gauge the enrichment performance of the program. Enrichment factor (EF)[52-54] is one of methods that is used to measure the virtual screening performance of the VinaLC docking program[40].

$$EF^{x\%} = \frac{Actives_{sampled}}{Actives_{total}} \frac{N_{total}}{N_{sampled}},$$

where $actives_{sampled}$ is the number of actives found at x% of the screened database, $actives_{total}$ is the number of total actives in the database, $N_{sampled}$ is the number of compounds at x% of database, and N_{total} is the number of total compounds in the database. The enrichment factor has several deficiencies because it largely depends on the composition of the data set and is not stable at low x%. Thus, in this study we used the average value of EF calculated from 38 targets in the DUD data set in order to eliminate the variability of data composition and reduce the uncertainty of the value at low x%. The EF values were calculated in two approaches, one uses the VinaLC score and other uses the MM/GBSA binding free energy.

Receiver Operating Characteristic (ROC) plot

The Receiver Operating Characteristic (ROC) plot is employed in this study for measuring virtual screening performance[55-57]. Figure 3 shows four different scenarios for ranking the actives and decoys according to either docking scores or MM/GBSA rescoring scores. In the first scenario, ideally, actives always rank better than the decoys, which yields ideal performance. In the second scenario, most of actives rank better than decoys, which yields good performance. The fraction of the actives and decoys are calculated when they are selected sequentially from the rank of their own collections (Figure 3). The ROC curve on the right with the x- and y-axis of the fractions for the

decoys and actives is plotted according the fractions combined in total sequence. The area under the curve (AUC) is greater than 0.5 in this scenario. In the third scenario, the actives and decoys rank equally, which yield random-selection performance. The ROC curve crawls along the diagonal line, and the AUC value is near 0.5. In the fourth and worse scenario, most decoys rank better than actives, which yields bad performance. The ROC curve is under the diagonal line, and the AUC values is less than 0.5. The ROC method can effectively differentiate two populations so that it can be applied to differentiate the actives against the non-active decoys and reveal the enrichment performance of the docking and rescoring methods.

Results and Discussion

The enrichment performance can be improved by MM/GBSA rescoring.

A total of 38 targets in the DUD data set with their actives and decoys in the DUD-E data set have been processed through the high throughput virtual screening pipeline. The plots of ROC curves for all 38 targets are shown in the supporting materials (Figure 1S). The ROC plots are arranged in sequence of 7 nuclear hormone receptors, 8 kinases, 3 serine proteases, 4 metalloenzymes, 2 folate enzymes, and 14 other protein targets[50]. The AUC values of the ROC curves were calculated and are shown in the barplots (Figure 4). Examining the enrichment performance from the perspective of enzyme classes shows that the enrichment performance between docking and free energy methods is dependent on the enzyme class. For folate enzymes and kinases, the MM-GBSA method significantly out performs the docking method in most cases. For the nuclear hormone receptors, the AUC values for docking and free energy rescoring are similar and have good enrichment performance. For the serine proteases, docking scores slightly out-performs the free energy rescoring. Both the docking and free energy method have poor enrichment performance against metalloenzymes. For the rest of the enzymes in the DUD data set, the MM/GBSA re-scoring method outperforms the docking enrichment in 11 out of 14 enzymes. The average AUC value from all 38 targets for Vina docking is 0.66, and the MM/GBSA rescoring method improves the AUC to 0.71, on average. Free energy rescoring differentiates the actives from the non-active decoys better than Vina docking. The average of the enrichment factors for all 38 targets at 0.5, 1, 2, 5, and 10% are shown in **Table 1**. The enrichment factors of the MM/GBSA rescoring at each of the percentages are consistently larger than their counterparts from Vina docking. Thus, the

early recovery of actives for MM/GBSA rescoring is consistently better than that of Vina docking. Many previous efforts using MM/GBSA rescoring[9,35-37,58-60] have been spent on pursuing better correlation between the calculated binding free energies and experimental pKi or pIC5, or finding structural accuracy of poses. Our study has systematically demonstrated with an array of various classes of proteins that free energy rescoring on average improves the enrichment performance.

Kinases

The DUD data set has 8 kinases as shown in Figure 4. One of the most significant performance differences between docking and free energy rescoring is with target hs90a. Target hs90a is the N terminal domain of heat shock protein (PDB ID: 1UYG), which contains a hydrophobic binding site[61]. Residues 104-111 adopt a helical conformation that is mainly hydrophobic in nature. In most cases, the aromatic rings of complexed ligands are stacked between the side chains of Phe138 and Leu107. Figure 5A shows the X-ray crystal ligand (8-(2,5-dimethoxy-benzyl)-2-fluoro-9h-purin-6-ylamine) in the active site of target hs90a as a reference. The ROC plot of hs90a shows the enrichment performance calculated from the Vina score is significantly worse than random selection with an AUC of 0.19, while the enrichment performance calculated from the MM/GBSA rescoring is better than random selection with an AUC of 0.56 (Figure 1S and Figure 4 hs90a). Only 17% of the ligands (both actives and decoys) select identical docking poses for the Vina score and MM/GBSA rescoring. About 37% of the decoys have Vina scores better (less) than -9.0 Kcal/mol while that for actives is only 12%. Overall, the decoys have unusually better Vina scores than the actives. Investigating the poses of decoys selected by docking and MM/GBSA rescoring, we found several types of docking poses that have consistent discrepancies, such as displacement from the active site, solvation

effects, and overestimation of interactions. In the first case, Vina-selected docking poses that prefer a different binding pocket. Decoy ZINC51634301, shown in Figure 5, clearly illustrates this problem. The docking pose has the ligand bound in a nearby tight pocket, completely displacing the ligand such that the ligand is no longer bound in the active site as defined by the crystal ligand, that is also shown in the same figure for comparison (Figure 5A). The free energy pose has the ligand bound in the original active site similar to the X-ray crystal ligand (Figure 5B). Obviously, the Vina docking method has picked the wrong binding mode, which can be attributed to the underestimation of the repulsion terms in the Vina scoring function. Judging from their ranking, the rank of Decoy ZINC51634301 from the docking pose is 265 out of a total of 5067 actives and decoys, and that for the MM/GBSA-selected pose is 1350 out of 5067, which shows that the decoy has been ranked significantly higher by the docking score. Similar cases can be found for Decoy ZINC47325294, ZINC44153979, ZINC10001939, and ZINC07613880. The second discrepancy between docking and free energy poses is related to solvation. Decoy ZINC12707283, which has two carbonyl groups (Figure 6), exemplifies this issue. The carbonyl groups of the docking pose are pointed into the hydrophobic active site while those of the MMGBSA-selected pose are exposed to bulk water. The Vina score ranks this decoy 1691 out of a total of 5067 actives and decoys while free energy rescoring ranks it merely 3751. Chemically, a carbonyl group is hydrophilic and should interact favorably with water. Thus, this type of discrepancy is likely due to an improved treatment of solvent by the GB model. Similar examples can be found in many other decoys, such as Decoy ZINC45513233, and ZINC16525498. The third difference is specific to aromatic ring-ring interactions energies. Decoy ZINC39856096 has seemingly

similar poses for both docking and free energy calculations. After energy minimization, the MM/GBSA rescoring selected structure deviates only slightly from the original Vina docking pose as shown in Figure 7. Structurally, there is no significant difference between the Vina and MM/GBSA methods. However, the interaction energy is significantly different. The MM/GBSA binding energy of Decoy ZINC39856096 puts the rank at the lowly position of 4162 out of a total of 5067 actives and decoys, while docking ranks the decoy 811 out of 5067. After careful examination of the complex structure, the Vina scoring function seems to overestimate the ring stacking effects between the side chains of Phe138 and Leu107, which may due to the overestimation of van der Waals effects (Figure 7). Similar to Decoy ZINC39856096, the docking score of Decoy ZINC39356214 ranks the ligand fairly high in the actives (456 out of 5067), while the free energy method ranks the ligand relatively low in the actives (3192 out of 5067). Also similar to Decoy ZINC39856096, Decoy ZINC39356214 has a phenyl ring but also has an additional quinazolinone ring (Figure 8). The MM/GBSA-selected pose prefers the quinazolinone ring stacking between the side chains of Phe138 and Leu107, while the docking pose has the phenyl ring stacking between them. The Vina scoring function prefers the homocyclic aromatic ring to the heterocyclic one for ring stacking. Overall, for target hs90a, the free energy method picks the correct pose as compared to Vina docking.

Another kinase, tyrosine-protein kinase (target src), has a highly conserved glutamic acid residue within a deep hydrophobic binding pocket that interacts with an amide/urea linker connecting the hydrophobic portions of the inhibitors[62]. This interaction is a key

feature to inhibitor binding. In general the inhibitors have different solvent properties compared to the decoy ligands. One feature of the inhibitors is the shifted aqueous solubility (LogSw). The average calculated LogSw[63] of the active ligands is -7.2 while the decoys have an average calculated LogSw of -6.6. Another difference between actives and decoys is the average calculated nonpolar solvent accessible surface area; active ligands are on average 319 $Å^2$ and the decoy ligands are 284 $Å^2$. The MM/GBSA rescoring calculation includes the solvent effect and captures the nonpolar interaction more accurately than the docking method so the performance is more representative of actual binding. The more accurate solvent and nonpolar interaction calculations in the MM/GBSA rescoring method lead to a better enrichment performance for most of the kinases, where the enrichment factor improves from 1.39, 1.39, 1.36, 1.36, 1.36 to 10.26, 9.74, 9.71, 7.28, 4.92 from Vina docking to MM/GBSA rescoring at 0.5%, 1%, 2%, 5%, 10% false positive rates, respectively. Another target egfr, epidermal growth factor receptor[64], has similar solvent accessible characteristics in the active site as that of target src, which makes predominantly hydrophobic interactions with ligands in its ATP binding site. Therefore, the enrichment factor of target egfr is improved by MM/GBSA rescoring as compared with docking in a same way as target src.

Metalloenzymes

The DUD data set includes metalloenzymes targets Angiotensin-converting enzyme (ace), Adenosine deaminase (ada), Catechol O-methyltransferase (comt), and Phosphodiesterase 5A (pde5a). The overall enrichment performance of metalloenzymes is barely better than random selection, which is largely due to the presence of the metal

ions in the active sites of the targets. For example, target ace (PDB ID 3BKL)[65] has a large active site much like a dumbbell (Figure 9A). The two large binding pockets are connected by a narrow channel, where the X-ray ligand binds to the channel by coordinating its two hydroxyl groups to the zinc atom. This configuration is very challenging for docking. As shown in the Figure 9B, the majority of the docking poses fall into either of the two large neighboring binding pockets. Most active ligands fail to coordinate with the zinc atom, which is crucial to binding in target ace. The ROC curves for target ace from both docking and free energy calculations (Figure 1S ace) show near random selection for active and decoy ligands. In the metalloenzymes, the metal ion plays an important role in the binding affinity and geometry by coordinating with the ligand. Unless this coordination is properly represented in the interaction energy, the ligand binding will be wrong. For most docking methods, including AutoDock Vina, the simplified scoring functions cannot characterize the coordination and, thus, lead to poor enrichment performance in all four metalloenzymes in the DUD data set.

Folate Enzymes

The glycinamide ribonucleotide (GAR) transformylase is a folate-dependent enzyme within the *de novo* purine biosynthetic pathway[66]. The binding site for the folate cofactor moiety in human GAR transformylase (PDB ID: 1NJS)[67], target pur2, consists of three portions: the pteridine binding cleft, the catalytic site, and the formyl transfer region (Figure 10A). In addition to the cofactor binding site, there is a substrate-binding site adjacent to the catalytic sites. The pteridine binding cleft provides negatively charged residues, Glu141, Asp142, and Asp144, and rich carbonyl groups from the backbone of the protein to form the hydrogen bonds with the active ligands. Many positively charged residues, His108, His121, Lys170, and His174, present in the substrate-binding site can

stabilize the negatively charged carboxyl groups of the active ligands. The same can also be found for Arg64 and Arg90 in the formyl transfer region. Thus, docking poses of the active ligands adopt two binding modes (Figure 10A). The majority of active ligands bind in the pteridine binding cleft, the catalytic site, and the substrate-binding site simultaneously. The rest of the ligands adopt the same binding mode as the folate cofactor moiety by binding in the pteridine binding cleft, the catalytic site, and the formyl transfer region simultaneously. The enrichment performance of target pur2 is excellent with AUC values of 0.922 for Vina docking and 0.996 for MM/GBSA rescoring. The molecular properties of decoy ligands were designed to be consistent with active ligands. Specifically, molecular weight, calculated LogP, H-bond donors and acceptors, number of rotatable bonds, and net molecular charge of decoys were chosen to match active compounds. Comparing the molecular properties of active and decoy ligands (**Table 2**), there is no significant difference in terms of molecular weight, hydrogen bond donors and acceptors, number of rings, and the fractional polar solvent accessible surface area. The difference is shown in the net molecular charge, lipophilicity (LogP), and especially the dipole. Looking at the distributions of LogP, active ligands span from -5 to 1 with a peak around -2.5, while decoys span a much wider range from -11 to 5 with a peak around 0.5. The distributions of LogP for active and decoy ligands have good overlap between -5 to -1. However, the distributions of the dipoles for active and decoy ligands are both narrow. The dipoles of active ligands mainly range from 40 to 100 D while that of decoys range from 0 to 40 D. There is no significant overlap between the two distributions of dipoles, which indicates the molecular dipole is the main factor that differentiates the active and decoy ligands. These results are expected because the pteridine binding cleft has

negatively charged residues and mostly binds with amine groups from ligands. On the other hand the substrate-binding site and formyl transfer region have rich positively charged residues and mostly attract the negatively charged carbonyl groups of ligands. The distances between the pteridine binding cleft and either the substrate-binding site or formyl transfer region are large. Thus, the active ligands must have a large dipole in order to bind with such a large bipolar binding sites. Both Vina docking and MM/GBSA rescoring methods capture the molecular dipole momentum very well and, thus, yield excellent enrichment performance. MM/GBSA rescoring has slightly better performance because the method has a more accurate calculation of the electrostatic interaction.

Target dyr, human dihydrofolate reductase, also has a folate binding site[68], which is a large bipolar binding site. Similar to target pur2, charged residues in the binding site have been shown to contribute significantly to electronic polarization of the folate ligand[69]. Scrutinizing the molecular properties for the ligand, the average LogP for active and decoy ligands are 0.97 and 1.16. The distributions of LogP for active and decoy ligands resemble each other closely. However, the average dipole momentum for active and decoy ligands for target dyr are 30.52 and 13.02 D. Judging from the results of both target pur2 and dyr, one can conclude that the dipole momentum of the ligand is crucial to binding the folate enzymes, and both Vina docking and MM/GBSA methods capture the dipole accurately. One difference between target dyr and pur2 is that distributions of dipoles of active and decoy ligands for target dyr are not well separated as those distributions for target pur2. Thus, the enrichment performance of the target pur2 is better than that of target dyr.

Serine Protease

Serine proteases cleave peptide bonds in proteins, where serine serves as the nucleophile in the reaction[70]. There are three binding pockets in the serine protease, S1, S2, and S3, as shown in Figure 11A. The catalytic triad, adjacent to the S2 binding pocket, consists of serine, histidine, and aspartic acid. Target try1, Trypsin I (PDB ID 2ayw), is a prominent member of serine protease family[71]. The diaminemethyl group of the crystal ligand forms hydrogen bonds with Asp189 in the S1 binding pocket (Figure 11B). Many actives of target try1 contain amines or diaminemethyl groups, which form hydrogen bonds with Asp189. Active CHEMBL327331 is one such active, containing a diaminemethyl group (Figure 11C), and docked into the S1 binding pocket as the crystal ligand (Figure 11D). However, the docking pose is not close enough for the ligand to form hydrogen bonds with Asp189, which causes the MM/GBSA rescoring method to drop this pose and select the pose with the diaminemethyl group exposed to the solvent. It is possible that the solvation model incorrectly over-stabilizes the diaminmethyl group solvation versus the stabilization of hydrogen bonds to the target. Therefore, the MM/GBSA rescoring method picks the pose with the wrong binding mode. About 13% of the actives of target try1 have a similar problem as Active CHEMBL327331. In these cases, the MM/GBSA rescoring results in the worse enrichment performance for target try1 as compared to docking. The AUC value of MM/GBSA is 0.63, which is smaller than that of Vina docking (0.78).

Nuclear Hormone Receptors

From the ROC plots (Figure 1S) and bar graphs of the AUC (Figure 4), the nuclear hormone receptors from the DUD data set have similar enrichment performance for docking and free energy rescoring. Peroxisome proliferator-activated receptors γ (target

pparg), has almost identical ROC curves for Vina docking and MM/GBSA rescoring. However, MM/GBSA rescoring chooses only ~18% of the same poses as Vina docking. Nevertheless, most of the poses selected by the MM/GBSA rescoring closely resemble the poses selected by Vina docking. The main differences between poses selected by the MM/GBSA rescoring and Vina docking are often in the areas that are exposed to solvent. The normalized distributions of the MM/GBSA score for actives and decoys have similar shapes as those from Vina score (Figure 12), which indicates that the MM/GBSA rescoring and Vina docking have similar enrichment performance for nuclear hormone receptors.

Other Enzymes

S-adenosylhomocysteine hydrolase (target sahh), catalyzes S-adenosylhomocysteine to adenosine and homocysteine in the hydrolytic direction and catalyzes the reverse reaction in the synthesis direction[72]. The active site is a highly charged, tight binding pocket. As mentioned in the previous section, the Vina scoring function underestimates the repulsion terms. When the decoys were docked into the tight active site, their scores are unusually high. About 40% of the decoys have Vina scores better (less) than -8.0 kcal/mol. MM/GBSA rescoring yields relatively accurate binding affinity. More than 82% of the decoys were determined to be lower affinity binder by MM/GBSA rescoring when using the criterion of -30 kcal/mol of MM/GBSA score. This also can be observed from the AUC values, where the AUC value of MM/GBSA (0.86) is much larger than that of Vina docking (0.56).

What is needed to achieve better enrichment performance?

MM/GBSA rescoring is dependent on the docking poses. Due to limited computer resources, only a few docking poses of the top resulting docking compounds can be

rescored by MM/GBSA. In this study, we choose only the top 20 docking poses for each ligand. The MM/GBSA rescoring, either calculated from a "single-point" energy minimization or averaged from MD trajectory, is often limited to explore the energy landscape of binding and usually constrained to the local minima. Thus, MM/GBSA cannot improve the accuracy of the calculated binding energy if the selected docking poses for rescoring have the wrong binding modes. Metalloenzymes employed in this study fall into this category; the docking poses have the wrong binding modes, and MM/GBSA rescoring cannot improve the enrichment performance.

The docking scoring function usually sacrifices accuracy for calculation speed. For example, the Vina scoring function uses pure empirical terms to speed up the docking calculations[73]. The tradeoff between the accuracy and calculation speed is subtle for MM/GBSA rescoring. The docking scoring function must pick poses with the correct binding mode for MM/GBSA rescoring. On average, the enrichment performance of Vina docking is good (as shown in the previous section and our previous study[40]). In some cases, the Vina scoring function underestimates the repulsion term, which can be readily corrected by MM/GBSA rescoring. MM/GBSA rescoring better accounts for the hydrophobic effect by estimating it from the solvent accessible surface. Such a hydrophobic effect arises from solute-imposed constraints on the organization of water that is part of entropic effects. The polarizable effect is also very important in the binding affinity calculation. Upon binding, the partial charges of atoms in both ligand and target will re-distribute[74,75], which can significantly affect binding. Unfortunately, both the Vina scoring function and MM/GBSA rescoring use fixed charge models, which do not

account for the polarizable effect. Quantum-mechanics-based or polarizable-force-field-based docking can capture the polarizable effect in this case[75-79].

Solvent effect plays an important role in binding affinity calculations. The standard Vina scoring function does not include solvation terms, which makes MM/GBSA rescoring particularly important to improve enrichment and obtain accurate binding affinity in cases where solvation plays a significant role. There are many examples in the DUD data set where ligands dock into the active site of solvent exposed targets. MM/GBSA is able to select the right poses with hydrophilic groups facing the solvent. Overall, MM/GBSA has a more accurate energy function than that of the Vina docking program. Most of the time MM/GBSA can pick the correct poses generated by Vina docking program.

How many poses are needed for accurate rescoring?

It is often the case that the best scoring docked pose produced by VinaLC does not correspond to the best ranked pose by MM-GBSA. Unfortunately, it can be computationally very expensive to rescore all poses produced by VinaLC. In order to evaluate the optimal number of docked poses to move forward to rescoring, we saved the top 20 docked poses of each complex for MM/GBSA rescoring. In practice, we performed ~14 millions MM/GBSA energyminimization calculations for the whole DUD dataset. With such a large amount of calculations, we were able to statistically determine the optimal number of docking poses that should be kept for MM/GBSA rescoring. In the first approach, we found the docking ranks (ranging from 1 to 20) of the MM/GBSA-selected poses for all 0.7 million complexes. The 0.7 million MM/GBSA-selected poses were then binned according to their docking ranks. The percentages of MM/GBSA-selected poses in each bin were calculated and a cumulative percentage plot is shown in Figure 13A. Docking-selected poses are always ranked 1. As seen in the Figure 13A,

only ~17% of complexes have identical docking- and MM/GBSA-selected poses. If the top 5 docking poses were kept, ~51% of complexes find the lowest MM/GBSA binding free energies among the top 20 poses. If the top 10 poses were kept, ~75% of complexes find the lowest values. ~90% of complexes find the lowest values if the top 15 poses were kept. Another approach is to calculate the average difference (i.e. error) between the minimal binding free energies of the top M and the top 20 docking poses, where M could be 1, 5, 10, 15, and 20.

$$Error(M) = \sum_{i=1}^{N} \left| \Delta G_{binding}^{min}(M) - \Delta G_{binding}^{min}(20) \right| \Big/ N$$

$\Delta G_{binding}^{min}(M)$ is the minimal binding free energy of top M poses calculated by MM/GBSA rescoring. N is the number of complexes. The Error(M) values are calculated for the actives and decoys separately and shown in Figure 13B. The trends of the lines for actives and decoys are similar. The errors at 1 are significantly large for both actives (9.5 kcal/mol) and decoys (9.8 kcal/mol). The lines become flat at the range from 5 to 10 poses. The errors of the free energy are 2.7 kcal/mol for actives and 2.8 kcal/mol for decoys at 5 poses. The errors of the free energy are 1.0 kcal/mol for actives and 1.1 kcal/mol for decoys at 10 poses. The errors of the free energy for both actives and decoys at 15 poses are 0.3 kcal/mol, which is fairly small as the average binding free energy for actives is about -30 kcal/mol. Error(M) values are the absolute errors and we also calculated the average error percentages by equation:

$$ErrorPercentage(M) = \sum_{i=1}^{N} \left| \left(\Delta G_{binding}^{min}(M) - \Delta G_{binding}^{min}(20) \right) / \Delta G_{binding}^{min}(20) \right| \Big/ N$$

The error percentages were plotted as shown in Figure 13C. When only one docking pose

is saved for rescoring, the error percentages are 33% and 35% for actives and decoys respectively. If the top 5 docking poses are rescored, the error percentages drop drastically to 10% and 12%. If the top 10 docking poses are rescored, the error percentages decrease to 4% for both actives and decoys. If the top 15 docking poses are rescored, the number drops to 1%. The lines of Error(M) and ErrorPercentage(M) drop significantly from 1 to 5 as compare to the line of cumulative percentage because it is quite often that the second or third (or more) best rescoring values are close to the best one. The errors will be small even if the best on are not selected. Picking the second best pose other than the best one may be acceptable ifthey are similar. Therefore, in terms of determining the optimal number of docking poses, Error(M) and ErrorPercentage(M) values are better than cumulative percentage values. Judged from all the above results, we suggest keeping at least the top 5 poses for the MM/GBSA rescoring. Ideally, we suggest keeping the top 10 docking poses to achieve good accuracy.

Does flexibility improve enrichment?

For a limited number of proteins, other sampling approaches were explored to assess if performance could be enriched other than through MM/GBSA rescoring. Flexible docking was performed for targets hs90a, ampc, pgh1, and nram due to their bad docking enrichment performance. Carefully selecting the flexible residues is crucial to the success of flexible docking[80]. In order to determine the flexible residues in the active sites, MD simulations of the receptors are carried out for several nanoseconds. The root mean square fluctuation (RMSF) values of the heavy atoms in side-chains for the residues within 7 Å of the active site center are calculated. The residues, which are not glycine, alanine, or proline and with RMSF values larger than 1 Å, were chosen as flexible residues. The enrichment performance of flexible docking for these four targets has

shown no significant improvement as compared to that of the rigid docking. Another approach was to select flexible residues by the B factors of the residues. After MD simulations, the flexible docking, using B factors to select residues, does not improve the enrichment performance either (Please see the supplemental materials figure S2 for the flexible docking results of target hs90a). Interestingly, increasing the exhaustiveness of the conformer search by four fold does not change the enrichment performance. The poor enrichment performance is most likely due to the failure of the scoring function to correctly characterize the binding between targets and ligands. For target hs90a, ampc, and nram, MM/GBSA rescoring does improve the enrichment performance. The lessons from our flexible docking exercise and other studies[7,80] are that selection of the flexible residues is important; however, flexible docking does not necessary yield better results than rigid docking.

Why parallel programming is essential for the docking and rescoring pipeline?

High throughput virtual screening of large databases is a very popular practice in computer-aided drug design. However, limited computational resources can cap the database size that can be screened in practice. The common practices of computer-aided drug design employed in pharmaceutical companies are still dominated by personal computers and mid-size clusters with hundred CPU cores. Supercomputers, although mature and popular in the field of molecular simulations and modeling, are seldom used in the *in-silico* drug design process. Applying high performance computing (HPC) toward drug design could be a game-changing strategy for pharmaceutical companies. For a typical example in the DUD data set, the time scales of a target of ~4K atoms and ~40K ligands running through the pipeline are shown in Figure 2. The process, if on a single

CPU, takes ~10 minutes to prepare the protein target, ~4 days to prepare the ligands, ~1 month to dock the ligands into the target, and ~10 years to rescore the top 20 docking poses of all ligands. In contrast, the same process takes ~1 hour on 100 CPUs to prepare the ligands, ~1 hour on 700 CPUs to carry out docking calculations, and ~5 hours on 15K CPUs to rescore the docking poses, totaling ~8 hours on HPC. HPC makes the screening of large databases practical and fits the fast pace of research and development in pharmaceutical companies. For example, a useful procedure is to screen 30 million compounds from the ZINC database against one therapeutic target using such a pipeline. One usually can down-select the 30 million compounds to several hundred thousands compounds, according to the rankings from docking calculations. The down-selected compounds then are rescored to further down-select for a drug lead. The whole procedure only takes days to finish on HPC but would be impractical to complete on a PC.

Conclusion

On average, the enrichment performance is improved by MM/GBSA rescoring. The average AUC value of ROC plots for MM/GBSA is larger than that of Vina docking. The early recovery of actives for MM/GBSA rescoring is consistently better than that of docking. However, the enrichment performance is target-dependent. MM/GBSA rescoring has better performance for folate enzyme, kinases, and several other enzymes.

MM/GBSA rescoring highly depends on the docking method to generate the poses with correct binding modes. As shown in the case of metalloenzymes, MM/GBSA rescoring cannot improve the accuracy if all the top ranked docking poses for rescoring have the wrong binding modes. The docking method has to sacrifice accuracy in the energy function for calculation speed. MM/GBSA has much more accurate energy functions to account for solvent effects, hydrophobic/entropic effects, electrostatic interaction, etc. However, the polarizable effect, one of the important factors in the binding calculation, is not taken into account for MM/GBSA and conventional docking methods. Flexible docking also does not necessary yield better results than the rigid docking. Determined by the statistical methods, we find the minimal number to keep the docking poses for MM/GBSA rescoring is 5 and the optimal number is 10.

To implement a docking and rescoring pipeline, such that better lead compounds can be discovered, requires parallel processing on HPC to screen millions of compounds. Applying HPC to *in-silico* drug design could be a game-changing strategy for pharmaceutical companies. The screening of large databases is not practical and may be impossible to complete on a PC, but is attainable within a day when implemented in

parallel on HPC machines. Thus, parallel docking and rescoring can impact daily decisions in drug design programs.

Acknowledgement

The authors thank Livermore Computing for the computer time and Laboratory Directed Research and Development for the funding (12-SI-004). This work was performed under the auspices of the United States Department of Energy by the Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Table

Table 1 Enrichment factor average from 38 DUD-E targets

Method	0.5%	1%	2%	5%	10%
Vina	10.78	7.48	5.54	4.05	2.81
MM/GBSA	12.94	8.99	7.70	5.30	3.68

Table 2 Molecular properties of active and decoy ligands from target pur2.

Molecular Property	Actives	Decoys
Molecular Weight	464±23	418±50
Hydrogen Bond Donors	4.3±0.6	3.1±1.2
Hydrogen Bond Acceptors	9.9±1.1	7.6±2.4
Number of rings	2.9±0.3	2.6±0.9
Fractional Polar SASA	0.48±0.05	0.41±0.10
Charge	-1.9±0.3	-0.9±0.6
LogP	-2.49±0.99	-0.04±2.14
Dipole Momentum (D)	61.7±32.9	19.5±12.1

Figures

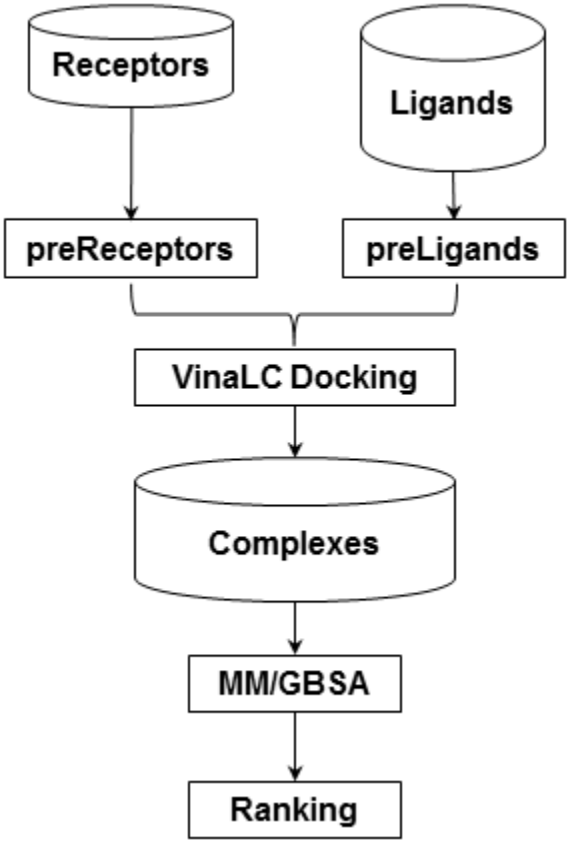


Figure 1. Flowchart of docking and rescoring.

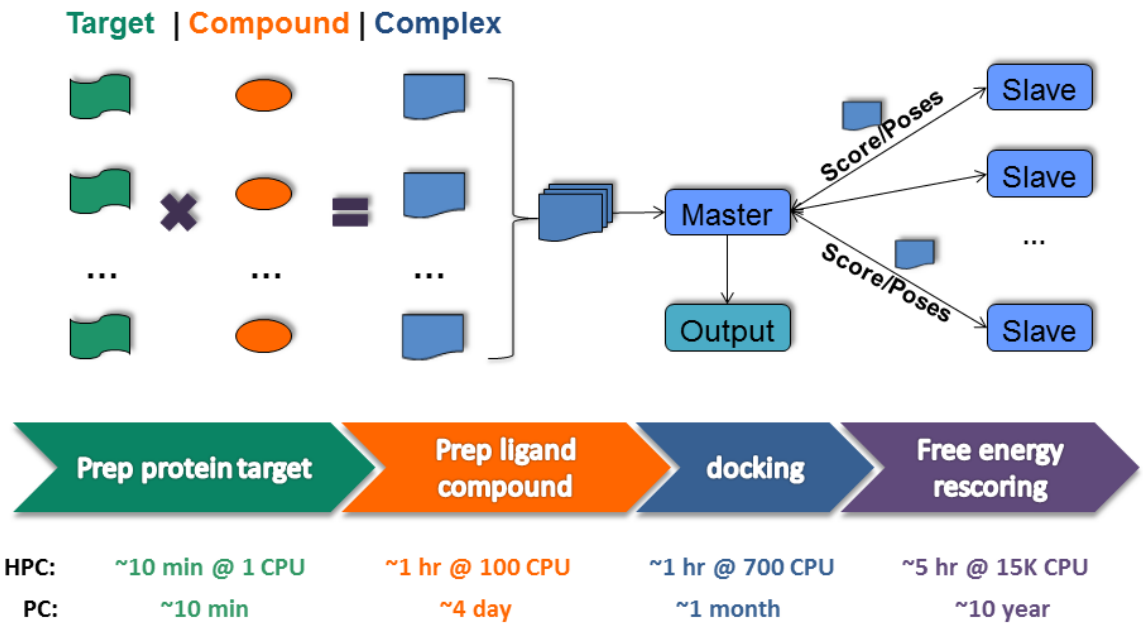


Figure 2. The pipeline and time scales for the docking and MM/GBSA rescoring calculations. The time scales are calculated from an example with a target of about 4K atoms and about 40K ligands. The top 20 docking poses are rescored using the MM/GBSA method.

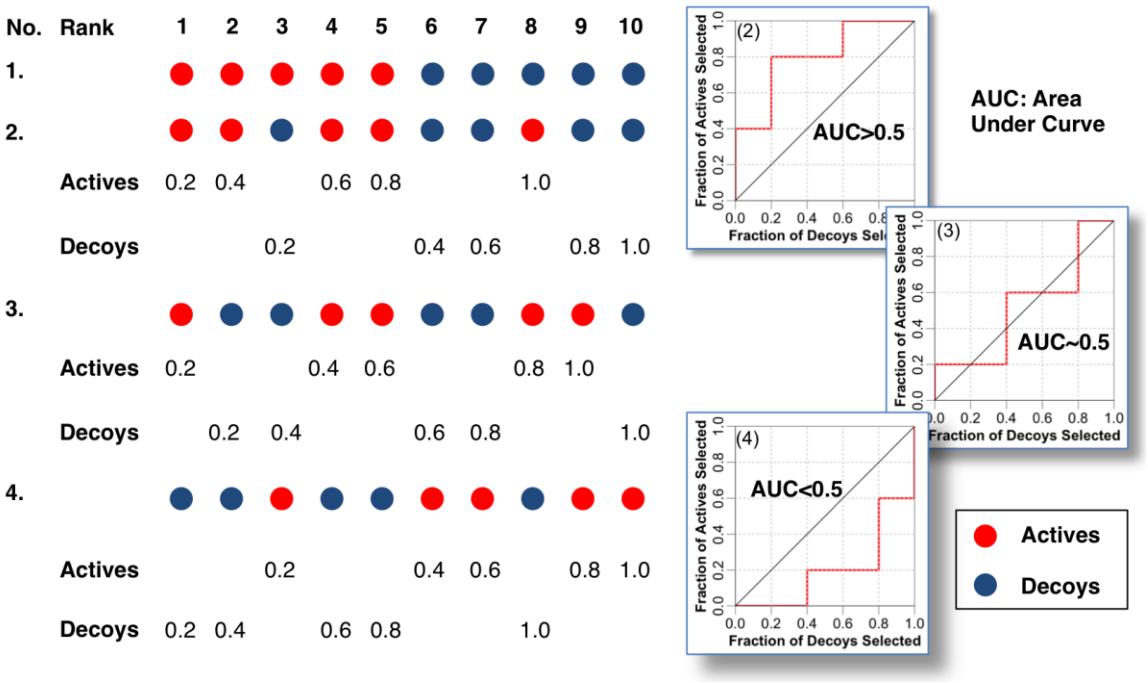


Figure 3. Scheme of Receiver Operating Characteristic (ROC) plots. The scheme depicts four different scenarios: (1) Ideal performance with actives always ranking better than decoys. (2) Good performance with most of actives ranking better than decoys. (3) Random performance with both actives and decoys ranking equally. (4) Bad performance with most of decoys ranking better than actives. The ROC plots of scenarios 2 to 4 are shown on the right.

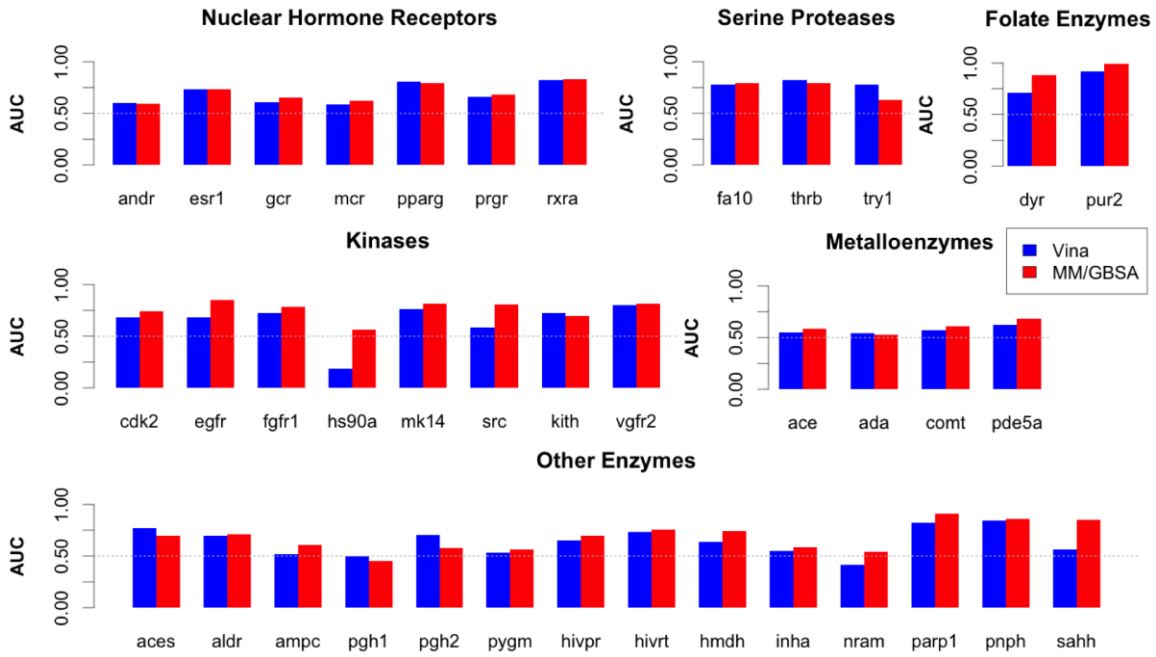


Figure 4. Bar graphs of the AUC values for the ROC curves of the DUD targets. The blue bars are the AUC for Vina docking and the red ones are for MM/GBSA rescoring. The graphs are arranged according to the enzyme type.

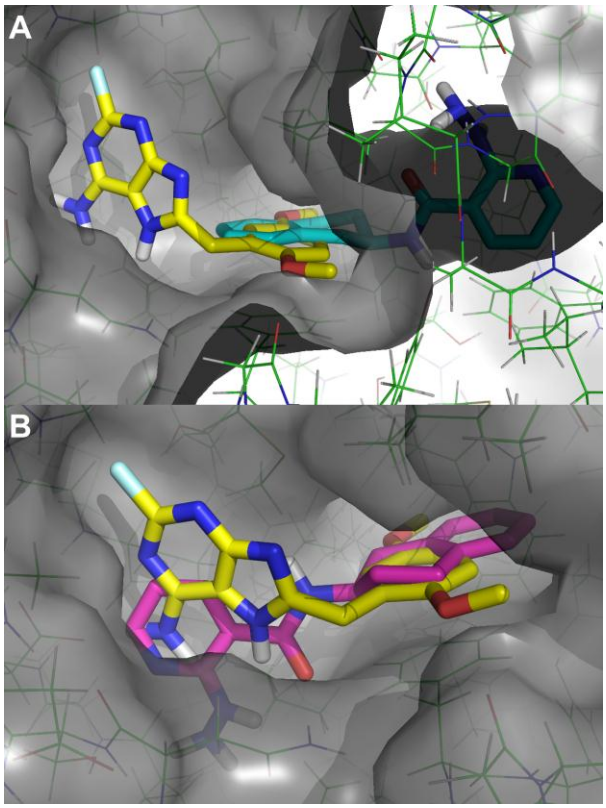


Figure 5. Poses of the Decoy ZINC51634301 in the active site of target hs90a: (A) docking pose (cyan) aligned with crystal ligand (yellow); (B) MM/GBSA rescoring pose (magenta) aligned with crystal ligand (yellow).

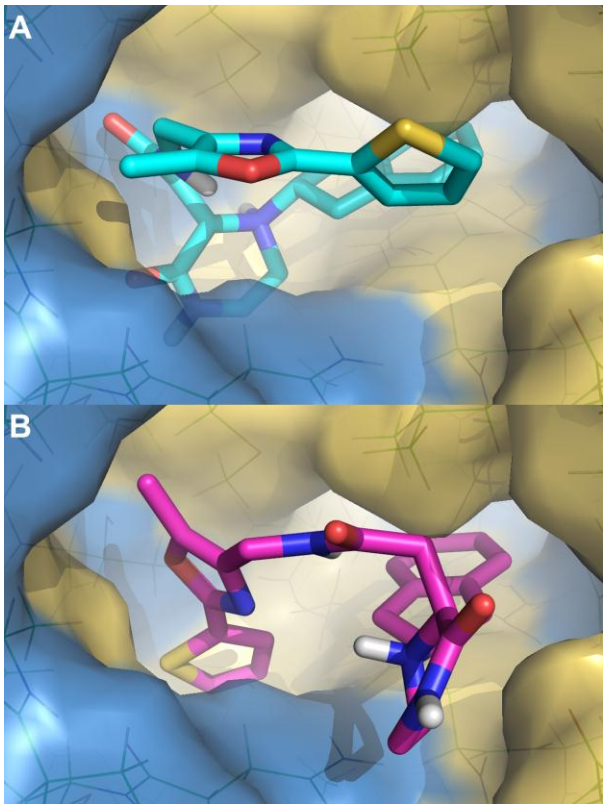


Figure 6. Poses of the Decoy ZINC12707283 in the active site of target hs90a: (A) docking pose; (B) MM/GBSA rescoring pose. The surfaces of protein are colored according to the hydrophobicity. The hydrophobicity decreases from yellow to blue .

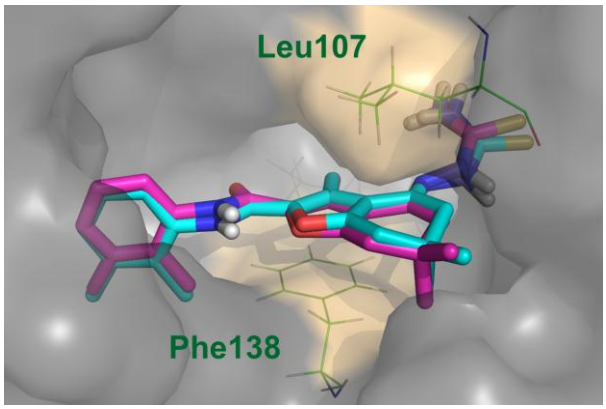


Figure 7. Poses of the Decoy ZINC39856096 in the active site of target hs90a. The carbon atoms of docking and MM/GBSA selected poses are colored in cyan and magenta, respectively.

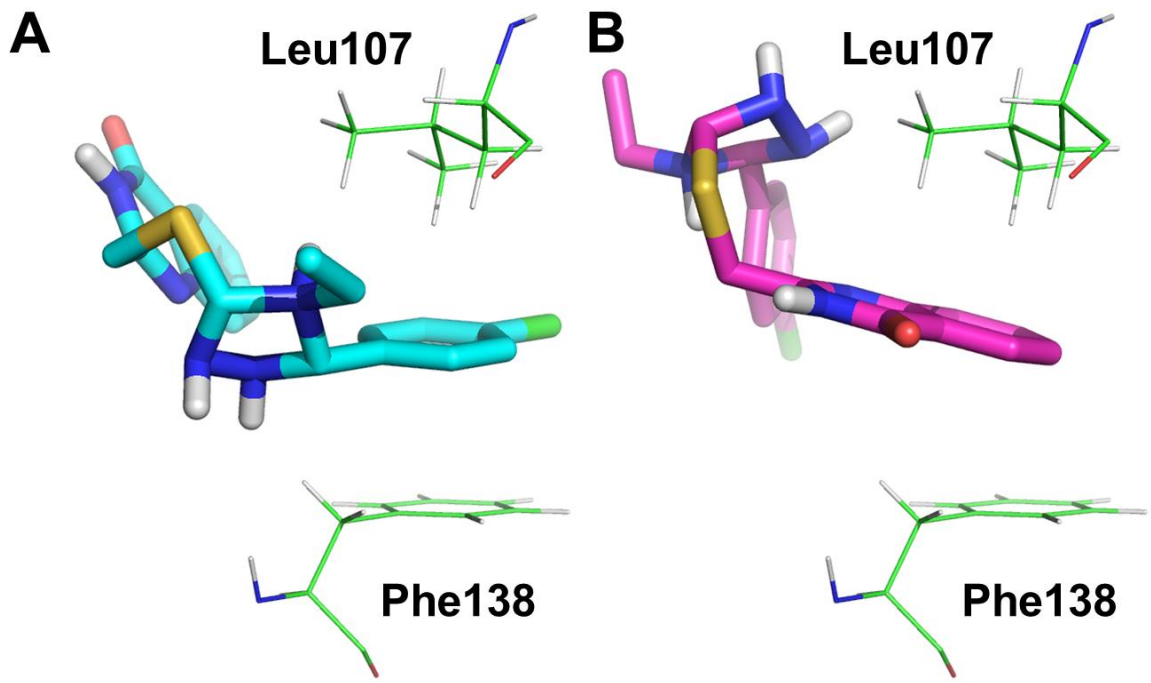


Figure 8. Poses of the Decoy ZINC39356214 in the active site of target hs90a: (A) Vina selected pose; (B) MM/GBSA selected pose.

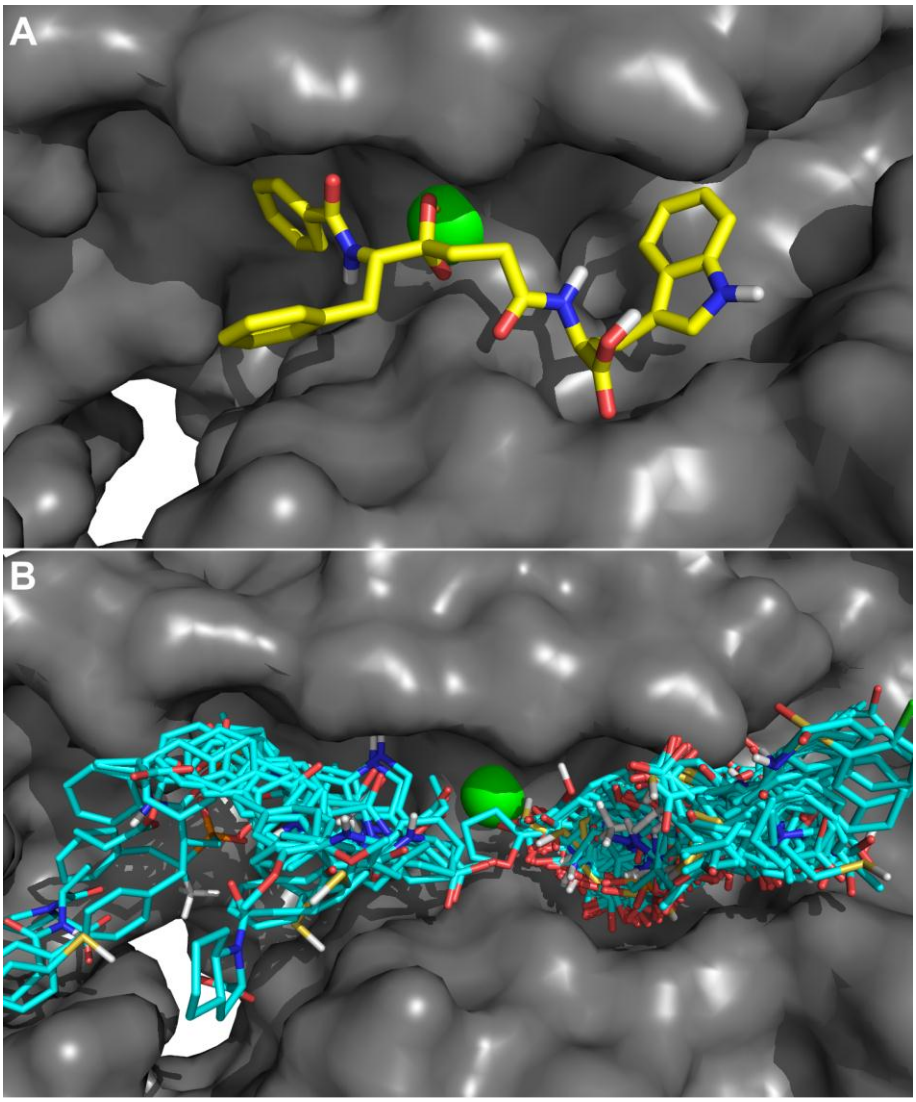


Figure 9. Cross-section of the target ace active site: (A) crystal ligand; (B) docking poses. The zinc atom is colored in green. Only 20% actives are shown by random selection in the figure to improve the visibility.

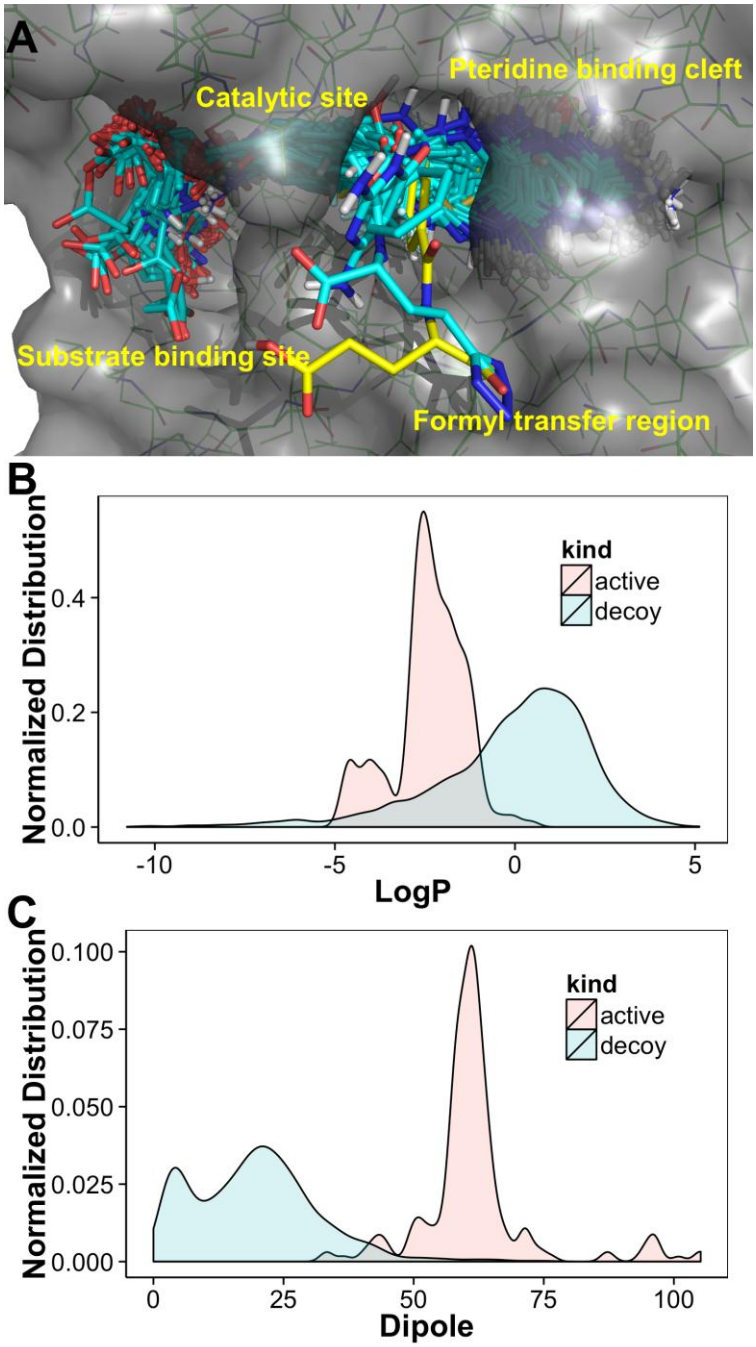


Figure 10. Target pur2. (A) Active site of target pur2 with 201 active ligands and an X-ray crystal ligand aligned in it. Carbon atoms of the active ligands are colored in cyan and those for X-ray crystal ligand are in yellow. (B) The normalized distribution of calculated LogP for active and decoy ligands. (C) The normalized distribution of calculated dipole for active and decoy ligands.

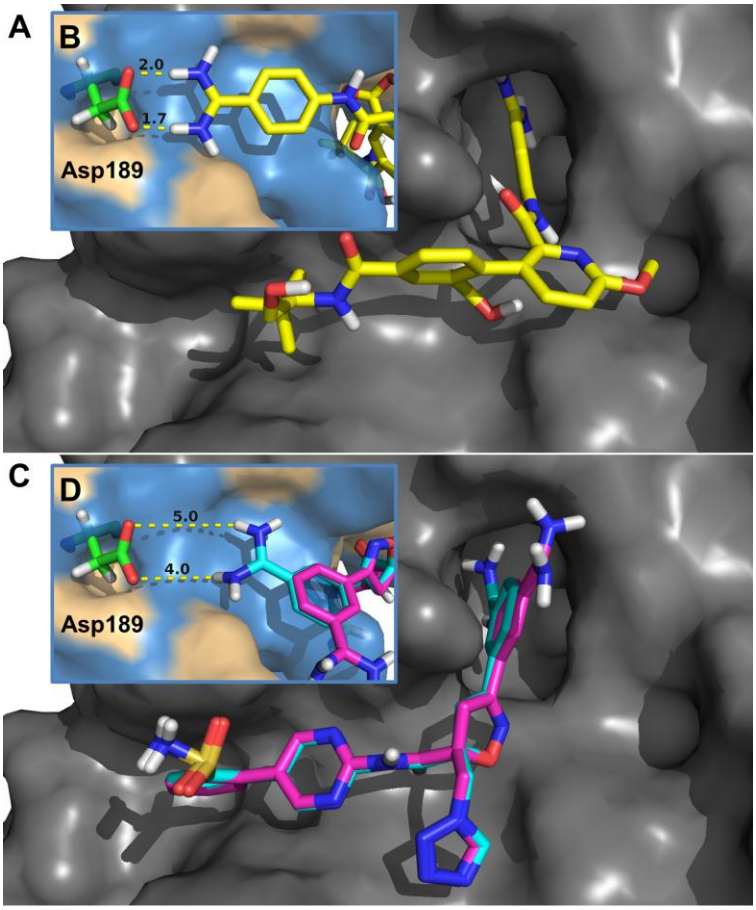


Figure 11. Target try1 with crystal ligand and Active CHEMBL327331: (A) crystal ligand; (B) crystal ligand in S1 binding pocket; (C) poses of Active CHEMBL327331 selected by docking and MM/GBSA rescoring. Carbon atoms of docking selected pose are colored in cyan and that for MM/GBSA rescoring are in magenta. (D) Active CHEMBL327331 in S1 binding pocket.

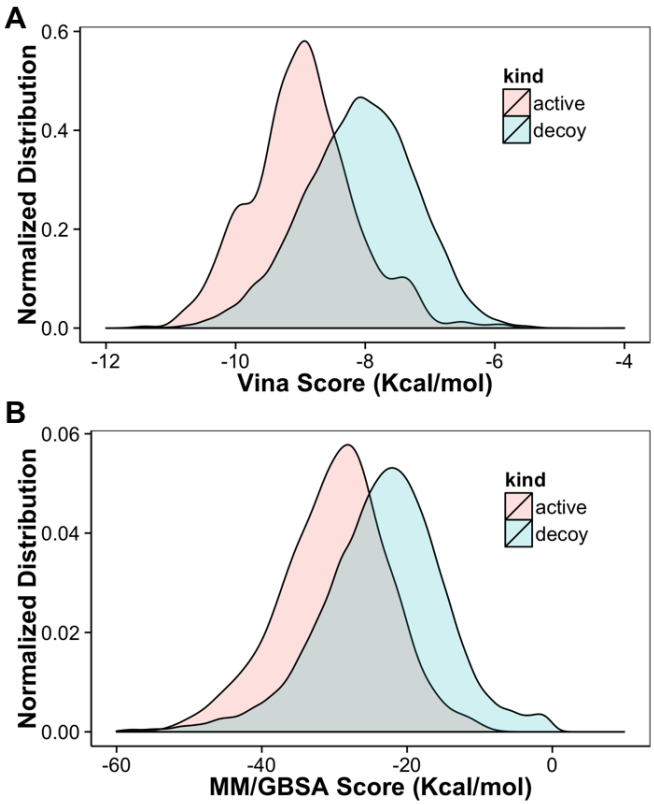


Figure 12. The normalized distribution of (A) Vina docking and (B) MM/GBSA rescoring score for active and decoy ligands.

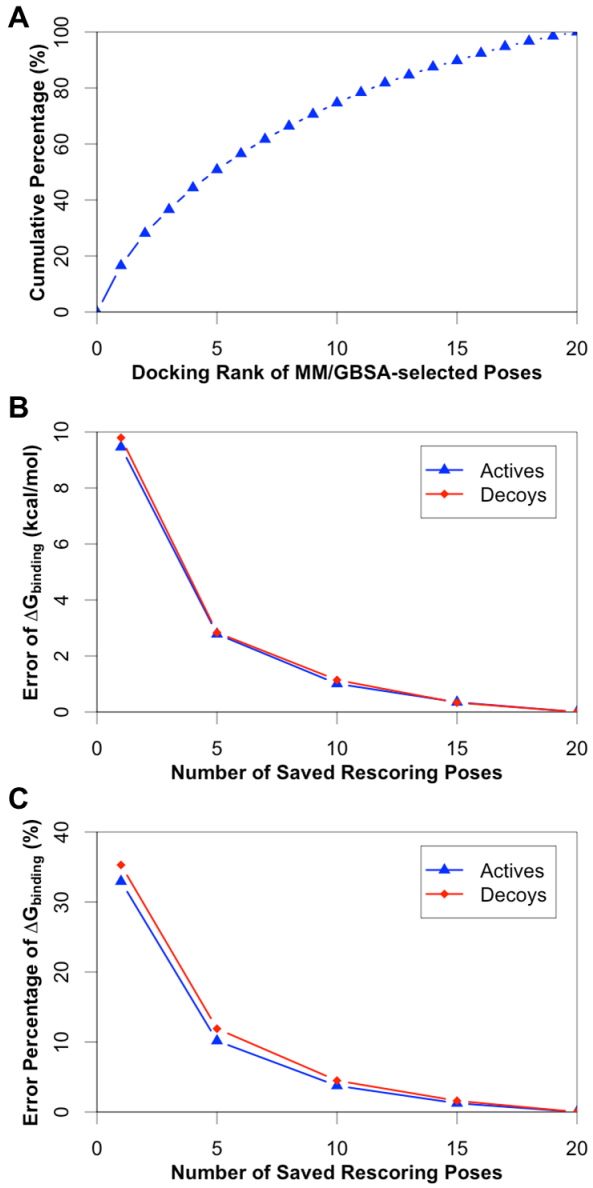


Figure 13. Determination of the optimal number of rescoring poses. (A) Cumulative percentage of MM/GBSA-selected poses versus their docking ranks. (B) Errors of $\Delta G_{binding}$ at different number of saved rescoring poses for actives and decoys respectively. (C) Error percentages of $\Delta G_{binding}$ at different number of saved rescoring poses for actives and decoys respectively.

Reference

1. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: Methods and applications. Nature Reviews Drug Discovery 3: 935-949.
2. Jorgensen WL (2004) The many roles of computation in drug discovery. Science 303: 1813-1818.
3. Smith RD, Dunbar JB, Ung PMU, Esposito EX, Yang CY, et al. (2011) CSAR Benchmark Exercise of 2010: Combined Evaluation Across All Submitted Scoring Functions. J Chem Inf Model 51: 2115-2131.
4. Halperin I, Ma BY, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins-Structure Function and Genetics 47: 409-443.
5. Wang RX, Lu YP, Wang SM (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46: 2287-2303.
6. Raha K, Peters MB, Wang B, Yu N, WollaCott AM, et al. (2007) The role of quantum mechanics in structure-based drug design. Drug Discovery Today 12: 725-731.
7. Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: Current status and future challenges. Proteins: Struct Funct Bioinf 65: 15-26.
8. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 153: S7-S26.
9. Rastelli G, Del Rio A, Degliesposti G, Sgobba M (2010) Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. J Comput Chem 31: 797-810.
10. Guimaraes CRW, Cardozo M (2008) MM-GB/SA rescoring of docking poses in structure-based lead optimization. J Chem Inf Model 48: 958-970.
11. Thompson DC, Humblet C, Joseph-McCarthy D (2008) Investigation of MM-PBSA rescoring of docking poses. J Chem Inf Model 48: 1081-1091.
12. Wong SE, Lightstone FC (2011) Accounting for water molecules in drug design. Expert Opin Drug Discov 6: 65-74.
13. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) COMPARISON OF SIMPLE POTENTIAL FUNCTIONS FOR SIMULATING LIQUID WATER. J Chem Phys 79: 926-935.
14. Roux B, Simonson T (1999) Implicit solvent models. Biophys Chem 78: 1-20.
15. Massova I, Kollman PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. Perspect Drug Discovery Des 18: 113-135.
16. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M (2005) Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 48: 4040-4048.
17. Gohlke H, Kiel C, Case DA (2003) Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaIGDS complexes. J Mol Biol 330: 891-913.
18. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J Comput Chem 26: 1668-1688.

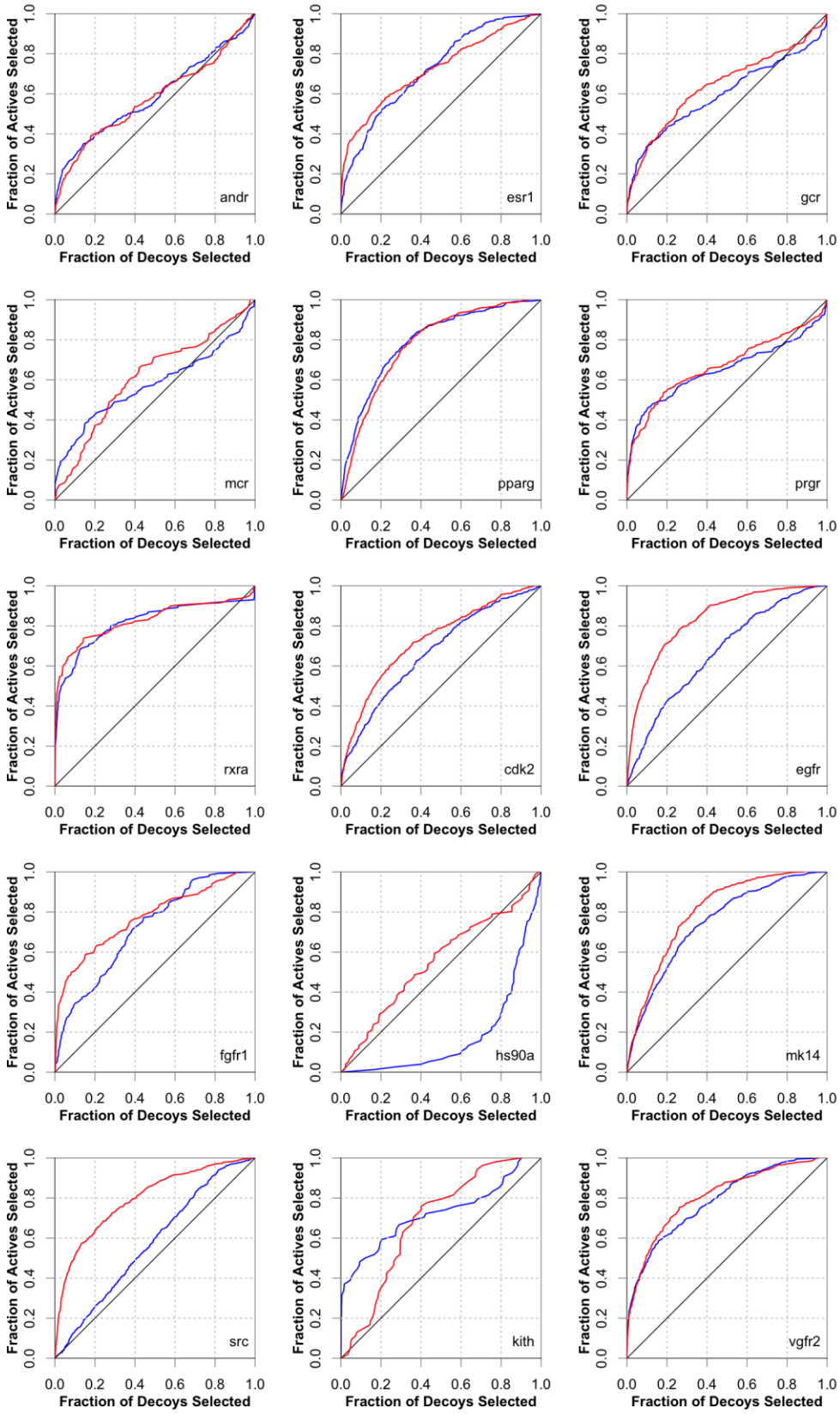
19. Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, et al. (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. J Chem Theory Comput 8: 3314-3321.
20. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: The Biomolecular Simulation Program. J Comput Chem 30: 1545-1614.
21. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4: 435-447.
22. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, et al. (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Acc Chem Res 33: 889-897.
23. Baker NA (2005) Improving implicit solvent simulations: a Poisson-centric view. Curr Opin Struct Biol 15: 137-143.
24. Feig M, Brooks CL (2004) Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol 14: 217-224.
25. Chen JH, Brooks CL, Khandogin J (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. Curr Opin Struct Biol 18: 140-148.
26. Lu BZ, Zhou YC, Holst MJ, McCammon JA (2008) Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. Commun Comput Phys 3: 973-1009.
27. Bruice TC (2006) Computational approaches: Reaction trajectories, structures, and atomic motions. Enzyme reactions and proficiency. Chem Rev 106: 3119-3139.
28. Mongan J, Case DA (2005) Biomolecular simulations at constant pH. Curr Opin Struct Biol 15: 157-163.
29. Chen JH, Im WP, Brooks CL (2006) Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. J Am Chem Soc 128: 3728-3736.
30. Kelly CP, Cramer CJ, Truhlar DG (2006) Adding explicit solvent molecules to continuum solvent calculations for the calculation of aqueous acid dissociation constants. J Phys Chem A 110: 2493-2499.
31. Shivakumar D, Deng YQ, Roux B (2009) Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. J Chem Theory Comput 5: 919-930.
32. Mongan J, Simmerling C, McCammon JA, Case DA, Onufriev A (2007) Generalized Born model with a simple, robust molecular volume correction. J Chem Theory Comput 3: 156-169.
33. Labute P (2008) The generalized Born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. J Comput Chem 29: 1693-1698.
34. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Struct Funct Bioinf 55: 383-394.
35. Hou TJ, Wang JM, Li YY, Wang W (2011) Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy

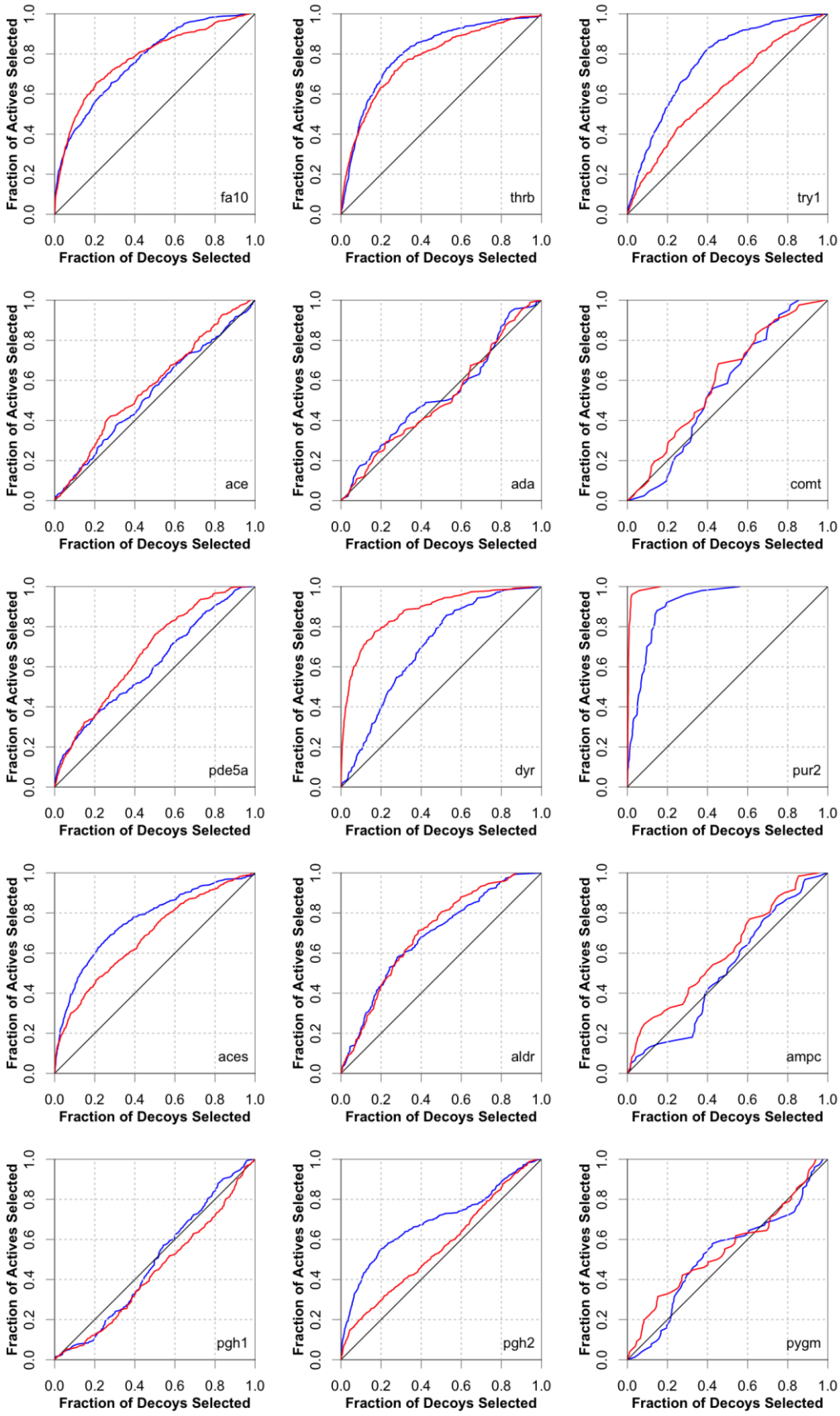
Calculations Based on Molecular Dynamics Simulations. J Chem Inf Model 51: 69-82.
36. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. J Comput Chem 32: 866-877.
37. Xu L, Sun H, Li Y, Wang J, Hou T (2013) Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. The Journal of Physical Chemistry B 10.1021/jp404160y.
38. Huang N, Kalyanaraman C, Irwin JJ, Jacobson MP (2006) Physics-based scoring of protein-ligand complexes: Enrichment of known inhibitors in large-scale virtual screening. J Chem Inf Model 46: 243-253.
39. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem 55: 6582-6594.
40. Zhang X, Wong SE, Lightstone FC (2013) Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines. J Comput Chem 34: 915-927.
41. Nilmeier JP, Kirshner DA, Wong SE, Lightstone FC (2013) Rapid Catalytic Template Searching as an Enzyme Function Prediction Procedure. PLoS One 8: e62535.
42. Kirshner DA, Nilmeier JP, Lightstone FC (2013) Catalytic site identification—a web server to identify catalytic site structural matches throughout PDB. Nucleic Acids Res 10.1093/nar/gkt403.
43. Richards FM (1977) AREAS, VOLUMES, PACKING, AND PROTEIN-STRUCTURE. Annu Rev Biophys Bioeng 6: 151-176.
44. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A GEOMETRIC APPROACH TO MACROMOLECULE-LIGAND INTERACTIONS. J Mol Biol 161: 269-288.
45. Ponder JW, Case DA (2003) Force fields for protein simulations. Protein Simulations 66: 27-+.
46. Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25: 1157-1174.
47. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graphics Modell 25: 247-260.
48. Jakalian A, Bush BL, Jack DB, Bayly CI (2000) Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. J Comput Chem 21: 132-146.
49. SLURM Simple Linux Utility for Resource Management: https://computing.llnl.gov/linux/slurm/.
50. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49: 6789-6801.
51. Maffucci I, Contini A (2013) Explicit ligand hydration shells improve the correlation between MM-PB/GBSA binding energies and experimental activities. J Chem Theory Comput 10.1021/ct400045d.

52. Pearlman DA, Charifson PS (2001) Improved scoring of ligand-protein interactions using OWFEG free energy grids. J Med Chem 44: 502-511.
53. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, et al. (2004) Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47: 1750-1759.
54. Cross JB, Thompson DC, Rai BK, Baber JC, Fan KY, et al. (2009) Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. J Chem Inf Model 49: 1455-1474.
55. Swets JA, Dawes RM, Monahan J (2000) Better decisions through science. Sci Am 283: 82-87.
56. Kellenberger E, Foata N, Rognan D (2008) Ranking targets in structure-based virtual screening of three-dimensional protein libraries: Methods and problems. J Chem Inf Model 48: 1014-1025.
57. Jain AN (2000) Morphological similarity: A 3D molecular similarity method correlated with protein-ligand recognition. J Comput Aided Mol Des 14: 199-213.
58. Hayes JM, Skamnaki VT, Archontis G, Lamprakis C, Sarrou J, et al. (2011) Kinetics, in silico docking, molecular dynamics, and MM-GBSA binding studies on prototype indirubins, KT5720, and staurosporine as phosphorylase kinase ATP-binding site inhibitors: The role of water molecules examined. Proteins: Struct Funct Bioinf 79: 703-719.
59. Greenidge PA, Kramer C, Mozziconacci JC, Wolf RM (2013) MM/GBSA Binding Energy Prediction on the PDBbind Data Set: Successes, Failures, and Directions for Further Improvement. J Chem Inf Model 53: 201-209.
60. Graves AP, Shivakumar DM, Boyce SE, Jacobson MP, Case DA, et al. (2008) Rescoring docking hit lists for model cavity sites: Predictions and experimental testing. J Mol Biol 377: 914-934.
61. Wright L, Barril X, Dymock B, Sheridan L, Surgenor A, et al. (2004) Structure-Activity Relationships in Purine-Based Inhibitor Binding to HSP90 Isoforms. Chem Biol 11: 775-785.
62. Dar AC, Lopez MS, Shokat KM (2008) Small Molecule Recognition of c-Src via the lmatinib-Binding Conformation. Chem Biol 15: 1015-1022.
63. Cheng AL, Merz KM (2003) Prediction of aqueous solubility of a diverse set of compounds using quantitative structure-property relationships. J Med Chem 46: 3572-3580.
64. Xu GZ, Abad MC, Connolly PJ, Neeper MP, Struble GT, et al. (2008) 4-amino-6-arylamino-pyrimidine-5-carbaldehyde hydrazones as potent ErbB-2/EGFR dual kinase inhibitors. Bioorg Med Chem Lett 18: 4615-4619.
65. Watermeyer JM, Kroger WL, O'Neill HG, Sewell BT, Sturrock ED (2008) Probing the basis of domain-dependent inhibition using novel ketone inhibitors of angiotensin-converting enzyme. Biochemistry (Mosc) 47: 5942-5950.
66. Inglese J, Johnson DL, Shiau A, Smith JM, Benkovic SJ (1990) SUBCLONING, CHARACTERIZATION, AND AFFINITY LABELING OF ESCHERICHIA-COLI GLYCINAMIDE RIBONUCLEOTIDE TRANSFORMYLASE. Biochemistry (Mosc) 29: 1436-1443.
67. Zhang Y, Desharnais J, Marsilje TH, Li CL, Hedrick MP, et al. (2003) Rational design, synthesis, evaluation, and crystal structure of a potent inhibitor of human

GAR tfase: 10-(trifluoroacetyl)-5,10-dideazaacyclic-5,6,7,8-tetrahydrofolic acid. Biochemistry (Mosc) 42: 6043-6056.
68. Cody V, Piraino J, Pace J, Li W, Gangjee A (2010) Preferential selection of isomer binding from chiral mixtures: alternate binding modes observed for the E and Z isomers of a series of 5-substituted 2,4-diaminofuro 2,3-d pyrimidines as ternary complexes with NADPH and human dihydrofolate reductase. Acta Crystallographica Section D-Biological Crystallography 66: 1271-1277.
69. Bajorath J, Kitson DH, Kraut J, Hagler AT (1991) The electrostatic potential of Escherichia coli dihydrofolate reductase. Proteins-Structure Function and Genetics 11: 1-12.
70. Hedstrom L (2002) Serine Protease Mechanism and Specificity. Chem Rev 102: 4501-4524.
71. Sherawat M, Kaur P, Perbandt M, Betzel C, Slusarchyk WA, et al. (2007) Structure of the complex of trypsin with a highly potent synthetic inhibitor at 0.97 angstrom resolution. Acta Crystallographica Section D-Biological Crystallography 63: 500-507.
72. Yang XD, Hu YB, Yin DH, Turner MA, Wang M, et al. (2003) Catalytic strategy of S-adenosyl-L-homocysteine hydrolase: Transition-state stabilization and the avoidance of abortive reactions. Biochemistry (Mosc) 42: 1900-1909.
73. Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31: 455-461.
74. Jiao D, Golubkov PA, Darden TA, Ren P (2008) Calculation of protein-ligand binding free energy by using a polarizable potential. Proc Natl Acad Sci U S A 105: 6290-6295.
75. Maple JR, Cao YX, Damm WG, Halgren TA, Kaminski GA, et al. (2005) A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. J Chem Theory Comput 1: 694-715.
76. Raha K, Merz KM (2005) Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. J Med Chem 48: 4558-4575.
77. Kantardjiev AA (2012) Quantum.Ligand.Dock: protein-ligand docking with quantum entanglement refinement on a GPU system. Nucleic Acids Res 40: W415-W422.
78. Gresh N (2006) Development, validation, and applications of anisotropic polarizable molecular mechanics to study ligand and drug-receptor interactions. Curr Pharm Des 12: 2121-2158.
79. Wang JM, Cieplak P, Li J, Hou TJ, Luo R, et al. (2011) Development of Polarizable Models for Molecular Mechanical Calculations I: Parameterization of Atomic Polarizability. J Phys Chem B 115: 3091-3099.
80. Lexa KW, Carlson HA (2012) Protein flexibility in docking and surface mapping. Q Rev Biophys 45: 301-343.

Supporting Materials




Fraction of Actives Selected
Fraction of Decoys Selected
0.0
0.2
0.4
0.6
0.8
1.0
fa10
thrb
try1
ace
ada
comt
pde5a
dyr
pur2
aces
aldr
ampc
pgh1
pgh2
pygm

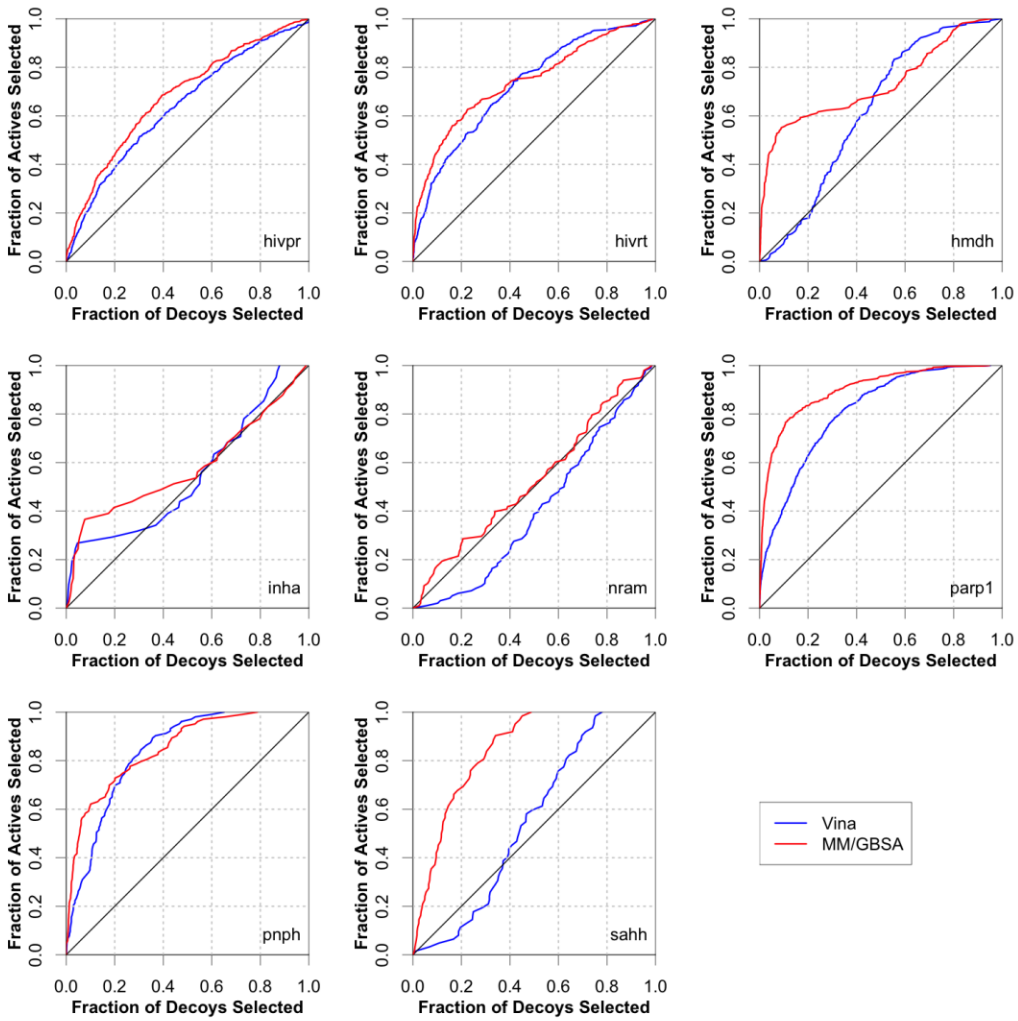


Figure S1. ROC plots of 38 DUD-E targets. The blue line is the ROC curve for the Vina score, and the red one is for the binding free energy calculated from the MM/GBSA method. The black diagonal line shows random selection.

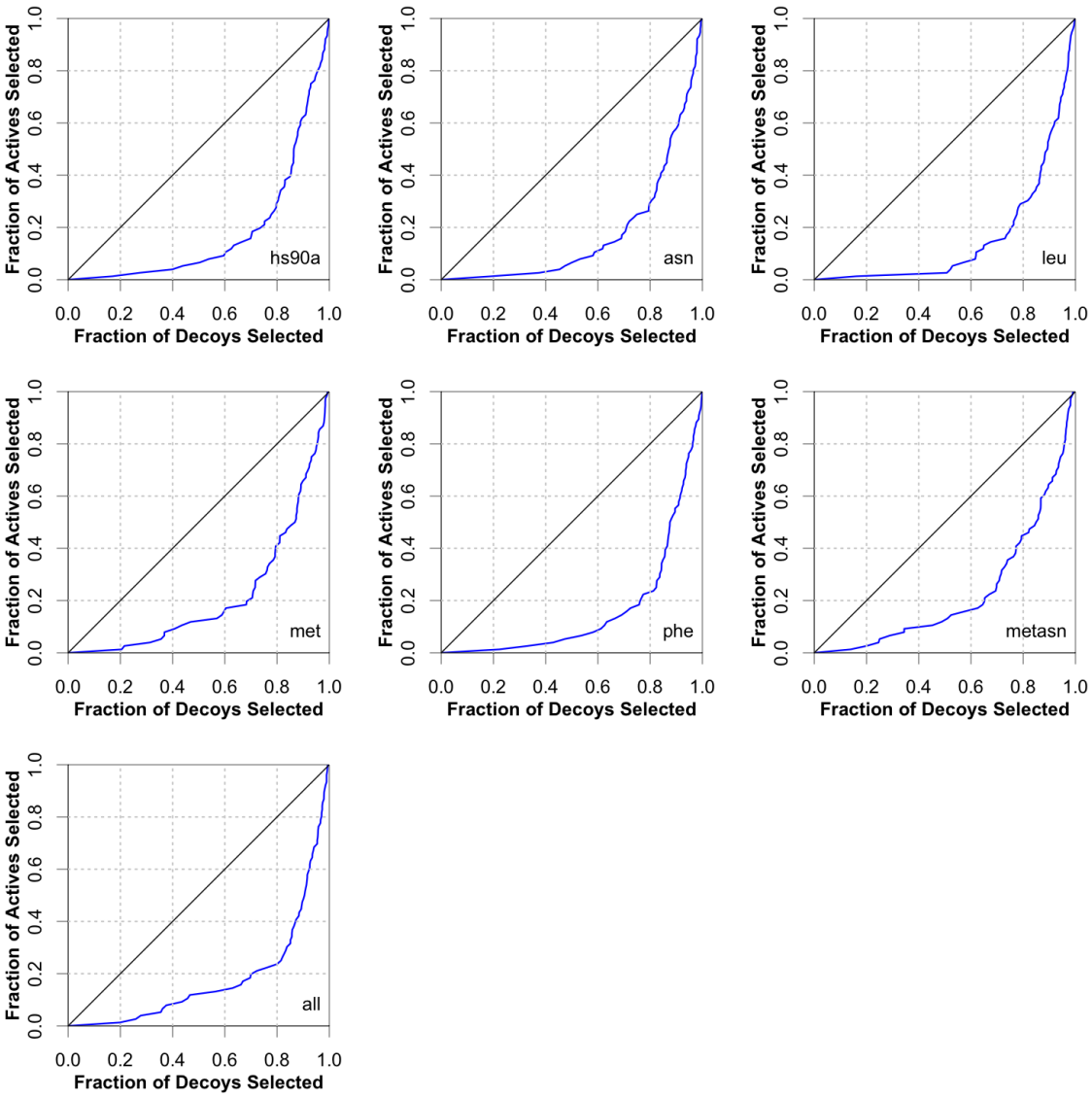


Figure S2. ROC plots of target hs90a flexible docking. 4 residues, Asn, Leu, Met, Phe, in the active site of target hs90a are selected for flexible docking according to RMSF values calculated from MD simulation. The first figure is from docking with 4 rigid residues. The next 4 figures allow one residue flexible at a time labeled by the names of flexible residues. The sixth figure allows two flexible residues, Met and Asn. The last figure allows all 4 flexible residues. There is no significant improvement by making residues flexible.